MSA101/MVE187 2017 Lecture 2

Petter Mostad

Chalmers University

August 31, 2017

Example: Learning about a proportion

- An experiment is performed n times. We assume there is a probability p for "success" each time, and that the outcomes are independent. Let X be the observed number of successes. We get $X \sim \text{Binomial}(n, p)$. Given $X = x$, what do we know about p?
- For a Bayesian analysis, we need a joint probability distribution (density) $\pi(X, p)$. We have defined $\pi(X \mid p)$ (the *likelihood*). We need to define $\pi(p)$ (the *prior*).
- Let us first try with the prior $p \sim \text{Uniform}[0, 1]$.
- The conditional model $\pi(p \mid X = x)$ (the *posterior* for p) can be computed with Bayes formula. We get

$$\pi(p \mid X = x) = \frac{\Gamma(n+2)}{\Gamma(x+1)\Gamma(n-x+1)} p^x (1-p)^{n-x}.$$

- We can recognize this as a Beta distribution: $p \mid X = x \sim \text{Beta}(x+1, n-x+1)$

Review of definition: The Beta distribution

θ has a Beta distribution on $[0, 1]$, with parameters α and β, if its density has the form

$$\pi(\theta \mid \alpha, \beta) = \frac{1}{\mathrm{B}(\alpha, \beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1}$$

where $\mathrm{B}(\alpha, \beta)$ is the Beta *function* defined by

$$\mathrm{B}(\alpha, \beta) = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$$

where $\Gamma(t)$ is the *Gamma function* defined by

$$\Gamma(t) = \int_0^\infty x^{t-1} e^{-x} \, dx$$

Recall that for positive integers, $\Gamma(n) = (n-1)! = 0 \cdot 1 \cdot \cdots \cdot (n-1)$. See for example Wikipedia for more properties of the Beta distribution, and the Beta and Gamma functions. We write $\pi(\theta \mid \alpha, \beta) = \mathrm{Beta}(\theta; \alpha, \beta)$ for the Beta density.

Using a Beta distribution as prior

- Assume the prior is $p \sim \text{Beta}(\alpha, \beta)$.
- The posterior becomes

$$p \mid (X = x) \sim \text{Beta}(\alpha + x, \beta + n - x)$$

- DEFINITION: Given a likelihood model $\pi(x \mid \theta)$. A *conjugate family of priors* to this likelihood is a parametric family of distributions so that if the prior for θ is in this family, the posterior $\theta \mid x$ is also in the family.

Using a discrete prior

- What if the prior for p is a discrete distribution, i.e., $\pi(p) = \sum_{i=1}^{k} I(p = p_i) q_i$?
- The conditional model is obtained with Bayes theorem:

$$P(p = p_i \mid x) = \frac{\pi(x \mid p = p_i) q_i}{\sum_{i=1}^{k} \pi(x \mid p = p_i) q_i} = \frac{p_i^x (1 - p_i)^{n-x} q_i}{\sum_{j=1}^{k} p_j^x (1 - p_j)^{n-x} q_j}.$$

- Computationally, you compute the vector of likelihoods, multiply termwise with the vector $(q_1, \dots, q_k)$ of prior probabilities, and normalize to 1.

Using discretization

- Assume you have ANY prior, with density $\pi(p)$ on $[0, 1]$. This density can be approximated, generally with reasonable accuracy, with a discrete distribution, a *discretization*.
- The corresponding posterior produced by discretization can be easily produced by computer: Compute the likelihood on a grid over p, compute the prior on the same grid, multiply, and normalize.
- NOTE: This works for ANY likelihood, as long as the parameter p has a prior distribution on a bounded set.

Discretizations useful in low dimensions

- The idea above can be extended to any model with 2 parameters, as long as they have a prior density on a bounded set. We come back with examples in the next lecture!
- This is an approximation. Accuracy will decrease dramatically when the number of (discretized) parameters increase beoynd 2 or 3 (why?). Thus discretization is rarely useful when there are more than 2-3 parameters.

Prediction

The Bayesian paradigm implies:

- The usefulness of a model lies in its ability to predict.
- We create a joint probability model for the parameters θ, the observed data x, and data we would like to predict x_{new}. Often on the form $\pi(\theta, x, x_{new}) = \pi(\theta)\pi(x \mid \theta)\pi(x_{new} \mid \theta)$.
- The distribution for x_{new} is given by conditioning on the observed x and marginalizing out θ:

$$\begin{aligned} \pi(x_{new} \mid x) &= \int_\theta \pi(\theta, x_{new} \mid x)\, d\theta = \int_\theta \pi(x_{new} \mid \theta, x)\pi(\theta \mid x)\, d\theta \\ &= \int_\theta \pi(x_{new} \mid \theta)\pi(\theta \mid x)\, d\theta \end{aligned}$$

  This is called the *posterior predictive distribution*.
- It is also possible to look at the predictive distribution for x before it has been observed. This is called the *prior predictive distribution*:

$$\pi(x) = \int_\theta \pi(x, \theta)\, d\theta = \int_\theta \pi(x \mid \theta)\pi(\theta)\, d\theta$$

Example: the Normal-Normal conjugacy

- Assume $\pi(x \mid \theta) = \text{Normal}(x; \theta, 1/\tau_0)$, where τ_0 is a known and fixed *precision*.
- Then $\pi(\theta \mid \mu, \tau) = \text{Normal}(\theta; \mu, 1/\tau)$, where τ is positive and μ has any real value, is a conjugate family.
- Specifically, we have the posterior

$$\pi(\theta \mid x) = \text{Normal}\left(\theta; \frac{\tau_0 x + \tau\mu}{\tau_0 + \tau}, \frac{1}{\tau_0 + \tau}\right)$$

- PROOF: Use completion of squares.

PROOF

$$
\begin{aligned}
\pi(\theta \mid x) &\propto_\theta \pi(x \mid \theta)\pi(\theta) \\
&\propto_\theta \exp\left(-\frac{\tau_0}{2}(x-\theta)^2\right)\exp\left(-\frac{\tau}{2}(\theta-\mu)^2\right) \\
&= \exp\left(-\frac{1}{2}\left[\tau_0 x^2 - 2\tau_0 x\theta + \tau_0\theta^2 + \tau\theta^2 - 2\tau\theta\mu + \tau\mu^2\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}\left[(\tau_0+\tau)\theta^2 - 2(\tau_0 x + \tau\mu)\theta\right]\right) \\
&\propto_\theta \exp\left(-\frac{1}{2}(\tau_0+\tau)\left(\theta - \frac{\tau_0 x + \tau\mu}{\tau_0+\tau}\right)^2\right) \\
&\propto_\theta \text{Normal}\left(\theta; \frac{\tau_0 x + \tau\mu}{\tau_0+\tau}, \frac{1}{\tau_0+\tau}\right)
\end{aligned}
$$